پیش‌بینی میزان فعالیت ترکیبات مولکولی با استفاده از انتخاب ویژگی نیمه‌نظارتی تُنک

راضیه شیخ‌پور

گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه اردکان، ایران

rsheikhpour@ardakan.ac.ir

چکیده

یکی از روش‌های محاسباتی برای پیش‌بینی فعالیت ترکیبات مولکول‌ها، رابطه کمی ساختار- فعالیت (QSAR) است. از چالش‌های موجود در مطالعات QSAR، صرف زمان و هزینه زیاد برای تعیین فعالیت تجربی ترکیبات، وجود تعداد زیاد ترکیبات با فعالیت ناشناخته و تعداد زیاد توصیف‌کننده‌های استخراج شده از ترکیبات است. انتخاب ویژگی نیمه‌نظارتی می‌تواند راه حل مناسبی برای پاسخگویی به چالش‌های مطالعات QSAR باشد که در فرآیند انتخاب ویژگی از داده‌های برچسب‌دار و بدون برچسب استفاده می‌کند. روش‌های انتخاب ویژگی نیمه‌نظارتی کلاسیک، ویژگی‌ها را جداگانه ارزیابی کرده و همبستگی میان ویژگی‌ها را در نظر نمی‌گیرد. در این مقاله، با استفاده از روش‌های انتخاب ویژگی نیمه‌نظارتی تُنک که همبستگی میان ویژگی‌ها را در نظر می‌گیرند، توصیف‌کننده‌های مناسب ترکیبات مولکولی تشخیص داده شده و با روش‌های انتخاب ویژگی نیمه‌نظارتی و غیرنظارتی کلاسیک و انتخاب ویژگی نظارتی تُنک مقایسه می‌شوند. نتایج آزمایش‌ها حاکی از برتری روش‌های انتخاب ویژگی نیمه‌نظارتی تُنک برای پیش‌بینی میزان فعالیت ترکیبات مولکولی است.

کلمات کلیدی: انتخاب ویژگی، نیمه‌نظارتی، تُنک، ترکیبات مولکولی، رابطه کمی ساختار- فعالیت

۱. مقدمه

پیش‌بینی فعالیت مولکول‌ها در چارچوب مطالعات رابطه کمی ساختار-فعالیت[1](QSAR) یکی از شاخه‌های علم کموانفورماتیک است که با استفاده از روش‌های نظری و محاسباتی می‌توانند فعالیت مولکول‌ها را پیش‌بینی نمایند. در این مطالعات، منظور اصلی کد کردن ساختار مولکول‌ها به‌صورت متغیرهای مستقل و یافتن ارتباط منطقی و کمی میان ساختار مولکول‌ها و فعالیت آنها است. [۳،۲،۱]. مطالعات QSAR، اساساً دارای پنج مرحله انتخاب داده‌های مناسب، محاسبه توصیف‌کننده‌ها[2]، انتخاب توصیف‌کننده‌های موثر، توسعه مدل و ارزیابی اعتبار مدل می‌باشند. برای مدل‌سازی رابطه بین

[1] Quantitative Structure-Activity Relationship
[2] Descriptors

ساختار ترکیبات و فعالیت آنها و انتخاب توصیف‌کننده‌های موثر، از تکنیک‌های مختلف یادگیری ماشین نظیر الگوریتم‌های رگرسیون[1] و روش‌های انتخاب ویژگی استفاده می‌شود.

یکی از چالش‌های یادگیری در مطالعات QSAR، تعداد زیاد توصیف‌کننده‌های استخراج شده برای هر ترکیب است که باید بتوان با استفاده از روش‌های انتخاب ویژگی، مناسب‌ترین توصیف‌کننده‌ها را برای دستیابی به مدل‌های QSAR با توانایی پیش‌بینی بالا انتخاب کرد. انتخاب ویژگی باعث اجتناب از بیش‌برازش[2] در هنگام ساخت مدل، بهبود کارایی و سادگی مدل می شود. علاوه‌بر این، انتخاب ویژگی نقش مهمی در درک و تجسم داده‌ها ایفا می کند.

از دیگر چالش‌های موجود در این مطالعات، وجود تعداد اندک ترکیبات با فعالیت تجربی شناخته شده است که با صرف زمان و هزینه زیادی توسط روش‌های تجربی به‌دست می‌آیند [۱]، درحالی‌که ترکیبات زیادی با فعالیت ناشناخته وجود دارند. در چنین مسائلی که ترکیبات کمی با فعالیت شناخته شده در دسترس هستند، روش‌های انتخاب ویژگی نظارتی ممکن است نتوانند توصیف‌کننده‌های مناسب را انتخاب کنند. برای حل این مشکل، در انتخاب ویژگی نیمه‌نظارتی علاوه‌بر داده‌های برچسب‌دار، از داده‌های بدون برچسب[3] نیز در فرآیند انتخاب ویژگی استفاده می‌شود [۴]. در روش‌های انتخاب ویژگی نیمه‌نظارتی، از برچسب داده‌های برچسب‌دار و اطلاعات توزیع و ساختار محلی داده‌های برچسب‌دار و بدون برچسب برای انتخاب ویژگی‌ها استفاده می‌شود.

روش‌های انتخاب ویژگی نیمه‌نظارتی کلاسیک نظیر امتیاز لاپلاسین نیمه‌نظارتی [۵]، اهمیت ویژگی‌ها را به صورت جداگانه ارزیابی می‌کنند و همبستگی[4] بین ویژگی‌ها را در فرآیند انتخاب ویژگی نادیده می‌گیرند. در راستای حل این مسئله، روش‌های انتخاب ویژگی تُنک[5] [۶–۱۰] ارائه شده‌اند تا همبستگی بین ویژگی‌ها را در هنگام انتخاب ویژگی در نظر بگیرند.

هدف این مقاله، استفاده از روش‌های انتخاب ویژگی نیمه‌نظارتی تُنک برای انتخاب توصیف‌کننده‌های موثر در ترکیبات مولکولی است تا بتوان از ترکیباتی که فعالیت آنها مشخص نشده است نیز در فرآیند انتخاب مناسب‌ترین توصیف‌کننده‌های برای بهبود عملکرد مدل‌های پیش‌ینی فعالیت ترکیبات استفاده کرد. برای این منظور از دو روش انتخاب ویژگی نیمه‌نظارتی تُنک استفاده شده و کارآیی آن بر روی دو مجموعه داده مولکولی با روش‌های انتخاب ویژگی نیمه‌نظارتی و غیرنظارتی کلاسیک و انتخاب ویژگی نظارتی تُنک مقایسه می‌شود. نتایج آزمایش‌ها، برتری روش‌های انتخاب ویژگی نیمه‌نظارتی تُنک در شناسایی توصیف‌کننده‌های مناسب برای پیش‌بینی میزان فعالیت ترکیبات مولکولی را نشان می‌دهند.

ادامه مقاله به صوت زیر سازماندهی شده است: در بخش ۲، روش‌های انتخاب ویژگی نیمه‌نظارتی تُنک برای انتخاب توصیف‌کننده‌های موثر شرح داده می‌شوند. در فصل ۳ نتایج آزمایش‌ها بیان شده و در بخش ۴، نتیجه‌گیری مقاله شرح داده می‌شود.

۲. روش‌های انتخاب ویژگی نیمه‌نظارتی تُنک

روش‌های انتخاب ویژگی نیمه‌نظارتی مبتنی بر مدل‌های تُنک، با استفاده از مدل‌های تُنک و داده‌های برچسب‌دار و بدون برچسب فرایند انتخاب ویژگی را انجام می‌دهند. هدف این روش‌ها، محاسبهٔ ماتریس انتقال $W \in R^{d\times c}$ بهینه برای انتخاب ویژگی است.

یک چارچوب کلی برای انتخاب ویژگی تُنک برای به‌دست آوردن W، کمینه نمودن تابع هدف زیر است [۱۱، ۱۲]:

[1] Regression
[2] Overfitting
[3] Unlabeled data
[4] Correlation
[5] Sparse

$$\min_{W} loss(W) + \lambda R(W) \qquad (۱)$$

که $loss(.)$ تابع زیان و $\lambda R(W)$ عبارت منظم‌سازی است که λ پارامتر منظم‌سازی است.

برای انتخاب ویژگی تُنک که همبستگی میان ویژگی‌ها را نیز در نظر بگیرد، به جای عبارت منظم‌سازی $\lambda R(W)$ در رابطه‌ی (۱)، معمولاً از منظم‌سازی نرم ماتریس-$l_{2,1}$ یا نرم ماتریس-$l_{2,1/2}$ استفاده می‌شود. نرم ماتریس-$l_{2,p}$ به‌صورت زیر تعریف می‌شود:

$$\|W\|_{2,p} = \left(\sum_{i=1}^{d} \|w^i\|_2^{1/p} \right)^{1/p} \qquad p \in (0,1] \qquad (۲)$$

هنگامی‌که $p=1$ باشد، نرم ماتریس-$l_{2,p}$ تبدیل به نرم-$l_{2,1}$ می‌شود.

فرض کنید مجموعه‌ای از n نمونه آموزشی $X = [x_1, \dots, x_l, x_{l+1}, \dots, x_n]^T$ شامل نمونه‌های برچسب‌دار و بدون برچسب وجود دارد که l تعداد داده‌های برچسب‌دار است. $x_i \in R^d \ (1 \le i \le n)$، iامین نمونه را مشخص می‌کند. همچنین فرض کنید $Y = [y_1, \dots, y_l, y_{l+1}, \dots, y_n]^T \in \{0,1\}^{n \times c}$ ماتریس برچسب داده‌های آموزشی باشد که c تعداد کلاس‌ها و $y_i \in R^c \ (1 \le i \le n)$، iامین بردار برچسب باشد. Y_{ij}، jامین داده‌ی y_i را مشخص کند، بنابراین اگر x_i در jامین کلاس باشد $Y_{ij}=1$ و در غیر این صورت $Y_{ij}=0$. اگر xi بدون برچسب باشد، y_i یک بردار با مقادیر صفر است.

منظم‌سازی منیفلد، یک روش معروف مبتنی بر لاپلاسین گراف است که الگوریتم‌های زیادی را تبدیل به الگوریتم‌های نیمه‌نظارتی نموده است. با به‌کار بردن منظم‌سازی منیفلد در تابع زیان رابطه‌ی (۱)، رابطه‌ی زیر به‌دست می‌آید.

$$\arg\min_{W,b} Tr(W^T X L X^T W) + \mu \left\| X_l^T W + 1_n b^T - Y_l \right\|_F^2 + \lambda \|W\|_{2,p} \qquad (۳)$$

که $b \in R^c$ یک عبارت بایاس[1] و $1_n \in R^n$ یک بردار ستونی است که همه‌ی n عنصر آن یک است. μ و λ پارامترهای منظم‌سازی هستند و L ماتریس لاپلاسین گراف است که بر اساس گراف k-نزدیک‌ترین همسایه به دست می‌آید.

ماتریس W به دست آمده با استفاده از رابطه‌ی (۳) تحت تأثیر برچسب داده‌های برچسب‌دار Y_l است. اگر بخواهیم برچسب همه‌ی داده‌های آموزشی را برای بهینه‌سازی W در نظر بگیریم، ماتریس برچسب‌های پیش‌بینی‌شده $F = [f_1, f_2, \dots, f_n]^T \in R^{n \times c}$ برای همه‌ی داده‌های آموزشی درنظر گرفته می‌شود که $f_i \in R^c \ (1 \le i \le n$ برچسب پیش‌بینی شده برای نمونه‌ی x_i است. ازآنجاکه F باید به برچسب‌های واقعی نزدیک باشد و بر روی گراف (ساختار منیفلد) هموار باشد، این ماتریس می‌تواند با کمینه نمودن تابع هدف زیر به دست آید:

$$\arg\min_{F} \sum_{l=1}^{c} \left[\frac{1}{2} \sum_{i,j=1}^{n} (F_{il} - F_{jl})^2 S_{ij} + \sum_{i=1}^{n} U_{ii} (F_{il} - Y_{il})^2 \right] \qquad (۴)$$

که F_{il}، lامین عنصر f_i یک ماتریس قطری $U \in R^{n \times n}$ است که ماتریس قانون تصمیم‌گیری نامیده می‌شود. در این ماتریس، اگر x_i داده‌ی برچسب‌دار باشد، مقادیر عناصر قطر اصلی بی‌نهایت است $(U_{ii}=\infty)$ و در غیر این صورت این مقادیر یک هستند $(U_{ii}=1)$. ماتریس قانون تصمیم‌گیری باعث سازگاری برچسب‌های پیش‌بینی شده توسط F با برچسب‌های واقعی Y می‌شود.

[1] Bias term

رابطه‌ی (۴) می‌تواند به‌صورت رابطه‌ی زیر نوشته شود:

$$\underset{F}{arg\,min}\, Tr(F^T LF) + Tr\Big((F-Y)^T U(F-Y)\Big) \quad (۵)$$

برای استفاده از داده‌های بدون برچسب، خطای پیش‌بینی می‌تواند با توجه به ماتریس پیش‌بینی برچسب کمینه شود. بنابراین تابع زیان در رابطه‌ی (۵) می‌تواند به‌صورت زیر باشد:

$$\underset{F,W,b}{\arg\min}\, Tr(F^T LF) + Tr\Big((F-Y)^T U(F-Y)\Big) + \mu\|X^T W + 1_n b^T - F\|_F^2 \quad (۶)$$

با تجمیع انتخاب ویژگی تُنک مبتنی بر نرم ماتریس-$l_{2,p}$ و لاپلاسین گراف مبتنی بر یادگیری نیمه‌نظارتی، تابع هدف زیر به‌دست می‌آید:

$$\underset{F,W,b}{\arg\min}\, Tr(F^T LF) + Tr\Big((F-Y)^T U(F-Y)\Big) + \mu\|X^T W + 1_n b^T - F\|_F^2 + \lambda\|W\|_{2,p} \quad (۷)$$

در تابع فوق، عبارت منظم‌سازی $\lambda\|W\|_{2,p}$ تضمین می‌کند که این مدل می‌تواند انتخاب ویژگی تُنک را انجام دهد. روش‌های انتخاب ویژگی ساختاری تُنک[1] ($SFSS$) [۱۱] و انتخاب ویژگی تُنک مبتنی بر لاپلاسین گراف[2] ($FSLG$) [۱۲] ، مبتنی بر رابطه (۷) به ترتیب از عبارت منظم‌سازی نُرم- $l_{2,1}$ و نُرم- $l_{2,1/2}$ برای انتخاب ویژگی‌ها استفاده می‌کنند.

۳. آزمایش‌ها

در این بخش، روش‌های انتخاب ویژگی نیمه‌نظارتی تُنک $SFSS$ و $FSLG$ برای انتخاب توصیف‌کننده‌های ترکیبات مولکولی ارزیابی می‌شوند.

۳. ۱. مجموعه داده‌ها

در این مقاله، دو مجموعه داده مولکولی استفاده می‌شود.

مجموعه داده مهارکننده‌های Rho کیناز[3] ($ROCK$): در این مجموعه داده [۱۳]، تعداد ۱۰۵ ترکیب $ROCK$ قرار دارند که ۳۵ ترکیب به عنوان داده‌های بدون برچسب، ۵۶ ترکیب به عنوان مجموعه آموزشی و ۱۴ ترکیب به عنوان مجموعه آزمون در نظر گرفته می‌شوند.

مجموعه داده مهارکننده‌های تیروزین-پروتئین کیناز FYN[4]: این مجموعه داده [۱۴]، شامل تعداد ۱۶۷ ترکیب تیروزین پروتئین کیناز FYN موجود در پایگاه bindingDB است که ۶۴ ترکیب به عنوان داده‌های بدون برچسب، ۸۰ ترکیب به عنوان داده آموزشی و ۲۳ ترکیب به عنوان داده آزمون درنظرگرفته می‌شوند.

[1] Structural Feature Selection with Sparsity
[2] Sparse Feature Selection based on Graph Laplacian
[3] Rho kinase (ROCK)
[4] Tyrosine-protein kinase FYN

۳. ۲. محاسبه توصیف‌کننده‌ها

برای محاسبه‌ی توصیف‌کننده‌ها، از ساختار دو بعدی و سه بعدی مولکول‌ها استفاده می‌شود. فایل ساختارهای دو بعدی و سه بعدی مولکول‌ها به نرم‌افزار PaDEL منتقل شده و برای هر ترکیب ۱۸۷۵ توصیف‌کننده شامل ۱۴۴۴ توصیف‌کننده‌ی یک بعدی و دو بعدی و ۴۳۱ توصیف‌کننده‌ی سه بعدی محاسبه می‌شود. مقدار توصیف‌کننده‌ها و فعالیت ترکیبات در بازه [۰/۰،۱/۹] نرمالیزه می‌شوند.

۳. ۳. روش‌های انتخاب ویژگی مقایسه‌شده

کارایی روش‌های انتخاب ویژگی نیمه‌نظارتی تُنک با تعدادی از روش‌های انتخاب ویژگی مقایسه می‌شود. این روش‌ها عبارتند از:

- انتخاب ویژگی مبتنی بر امتیاز لاپلاسین نیمه‌نظارتی[1] برای مسائل رگرسیون ($SSLS$) [۵]: روش امتیاز لاپلاسین نیمه‌نظارتی، با استفاده از اطلاعات برچسب داده‌های برچسب‌دار و ساختار محلی داده‌ها ویژگی‌ها را ارزیابی می‌کند. این روش، همبستگی بین ویژگی‌ها را در هنگام انتخاب ویژگی در نظر نمی‌گیرد و آنها را یکی یکی انتخاب می‌کند.
- انتخاب ویژگی با استفاده از کمینه‌سازی نُرم-$l_{2,1}$ ($FSNM$)[2] [۱۵]: روش $FSNM$، یک روش انتخاب ویژگی تُنک مبتنی بر نُرم- $l_{2,1}$ است که به‌صورت نظارتی عمل می‌کند. این روش، کمینه‌سازی نُرم- $l_{2,1}$ را بر روی تابع زیان و عبارت منظم‌سازی اعمال می‌کند.
- انتخاب ویژگی تُنک مبتنی بر کمینه‌سازی نُرم-$l_{2,1/2}$ ($SFSN$)[3] [۱۶]: روش $SFSN$، یک روش انتخاب ویژگی نظارتی تُنک و پایدار است که کمینه‌سازی نُرم-$l_{2,1/2}$ را بر روی تابع زیان و عبارت منظم‌سازی به‌کار می‌گیرد.
- انتخاب ویژگی مبتنی بر امتیاز لاپلاسین (LS) [۱۷]: امتیاز لاپلاسین، یک روش انتخاب ویژگی غیرنظارتی است که ویژگی‌ها را از طریق توانایی آنها در حفظ ساختار محلی داده‌ها ارزیابی می‌کند. این روش، همبستگی بین ویژگی‌ها را در هنگام انتخاب ویژگی نادیده گرفته و آنها را یکی یکی ارزیابی می‌کند.

۳. ۴. مدل‌سازی و ارزیابی اعتبار مدل

پس از انتخاب مناسب‌ترین توصیف‌کننده‌ها، برای برقراری ارتباط بین توصیف‌کننده‌های منتخب و فعالیت ترکیبات مولکولی با استفاده از ترکیبات برچسب‌دار، از مدل رگرسیون ناپارامتری مبتنی بر هسته گاوسی استفاده می‌شود و اعتبار مدل با استفاده از معیارهای ضریب همبستگی بین مقادیر تجربی و پیش‌بینی‌شده (R^2)، ضریب همبستگی تطابق (CCC)، ریشه میانگین مربعات خطا ($RMSE$) و میانگین خطای مطلق (MAE) ارزیابی می‌شود. این معیارها به صورت زیر محاسبه می‌شوند.

$$R^2 = \frac{\left[\sum_{i=1}^{n}\left((y_i - \bar{Y}_{exp}) \times (\hat{y}_i - \bar{Y}_{pred})\right)\right]^2}{\sum_{i=1}^{n}(y_i - \bar{Y}_{exp})^2 \times \sum_{i=1}^{n}(\hat{y}_i - \bar{Y}_{pred})^2} \quad (۸)$$

[1] Semi-supervised Laplacian score
[2] Feature Selection via Joint $l_{2,1}$-Norms Minimization
[3] Sparse Feature Selection based on $l_{2,1/2}$-Norms Minimization

$$CCC = \frac{2\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \hat{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - \hat{y})^2 + n(\bar{y} - \hat{y})^2} \quad (۹)$$

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \quad (۱۰)$$

$$RMSE = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (۱۱)$$

در رابطه‌های فوق، n تعداد ترکیبات، y_i مقدار تجربی فعالیت iامین ترکیب و $\hat{y}_i$ مقدار پیش‌بینی‌شده فعالیت iامین ترکیب را نشان می‌دهد. $\bar{Y}_{exp}$ و $\bar{Y}_{pred}$ به‌ترتیب میانگین مقادیر تجربی و پیش‌بینی‌شده فعالیت‌ها را بیان می‌کنند.

جداول ۱ و ۲ نتایج روش‌های مختلف انتخاب ویژگی را براساس معیارهای ذکر شده به ترتیب بر روی مجموعه داده‌های *ROCK* و *FYN* نشان می‌دهند.

جدول ۱- نتایج روش‌های مختلف انتخاب ویژگی براساس معیارهای ارزیابی مختلف بر روی مجموعه داده *ROCK*

روش انتخاب ویژگی	تعداد ویژگی‌های انتخاب شده	R^2	*CCC*	*RMSE*	*MAE*
SFSS	۱۱	۰/۸۸۳۲	۰/۹۰۹۲	۰/۳۱۴۷	۰/۲۲۵۳
FSLG	۸	۰/۸۲۷۷	۰/۸۸۴۴	۰/۳۷۴۷	۰/۲۸۱۳
SSLS	۱۱	۰/۷۶۶۸	۰/۷۷۱۲	۰/۴۶۲۲	۰/۳۷۸۵
FSNM	۶	۰/۴۶۹۳	۰/۵۶۳۴	۰/۶۱۳۱	۰/۴۸۵۲
SFSN	۱۱	۰/۴۶۰۴	۰/۵۹۰۰	۰/۶۰۹۴	۰/۴۹۶۵
LS	۱۱	۰/۵۱۵۷	۰/۷۰۴۲	۰/۶۰۲۹	۰/۴۷۸۱

جدول ۲: نتایج روش‌های مختلف انتخاب ویژگی براساس معیارهای ارزیابی مختلف بر روی مجموعه داده *FYN*

روش انتخاب ویژگی	تعداد ویژگی‌های انتخاب شده	R^2	*CCC*	*RMSE*	*MAE*
SFSS	۱۲	۰/۶۸۰۵	۰/۷۸۷۰	۰/۷۹۰۹	۰/۵۹۰۷
FSLG	۱۶	۰/۶۷۶۰	۰/۸۲۰۹	۰/۸۱۰۱	۰/۶۵۷۹
SSLS	۸	۰/۵۱۳۰	۰/۶۵۰۹	۰/۹۷۷۹	۰/۷۴۸۶
FSNM	۱۶	۰/۴۰۲۳	۰/۵۰۶۷	۱/۰۹۰۰	۰/۷۱۰۳
SFSN	۱۲	۰/۳۵۶۴	۰/۵۱۴۷	۱/۱۰۹۰	۰/۸۹۷۸
LS	۱۲	۰/۴۴۸۳	۰/۴۹۴۹	۱/۰۸۷۸	۰/۸۹۹۹

همان‌گونه که در جداول ۱ و ۲ مشخص است، روش‌های انتخاب ویژگی نیمه‌نظارتی تُنک دارای ضریب همبستگی بالاتری بر اساس معیارهای R^2 و CCC و خطای پایین‌تری بر اساس معیارهای $RMSE$ و MAE هستند.

نتایج جداول ۱ و ۲ نشان می‌دهند که روش‌های انتخاب ویژگی نیمه‌نظارتی تُنک در مقایسه با روش انتخاب ویژگی نیمه‌نظارتی و غیرنظارتی کلاسیک و روش‌های انتخاب ویژگی نظارتی تُنک کارایی بهتری دارند که این امر به دلیل استفاده از داده‌های بدون برچسب به همراه داده‌های برچسب‌دار و در نظر گرفتن همبستگی بین ویژگی‌ها است.

۴. نتیجه‌گیری

رابطه کمی ساختار- فعالیت (QSAR) یک روش محاسباتی است که فعالیت بیولوژیکی یا بیوشیمیایی ترکیبات مولکولی را به طور کمی به ساختار آنها مربوط می‌سازد. یکی از چالش‌های موجود در این مطالعات، تعداد زیاد توصیف‌کننده‌های استخراج شده از ساختارهای ترکیبات مولکولی است که بسیاری از آنها حاوی اطلاعات مفیدی نیستند. از طرفی دیگر، ترکیبات با فعالیت تجربی شناخته‌شده، با صرف زمان و هزینه زیادی به‌دست می‌آیند، از این‌رو تعداد آنها اندک است درحالی‌که ترکیبات زیادی با فعالیت‌های ناشناخته وجود دارند. برای پاسخگویی به چالش‌های مطالعات QSAR، روش‌های انتخاب ویژگی نیمه‌نظارتی می‌تواند مورد استفاده قرار گیرند. در روش‌های انتخاب ویژگی نیمه‌نظارتی کلاسیک، اهمیت ویژگی‌ها جداگانه ارزیابی می‌شوند. در این مقاله، از روش‌های انتخاب ویژگی نیمه‌نظارتی تُنک مبتنی بر نُرم-$l_{2,1}$ و نُرم-$l_{2,1/2}$ برای انتخاب توصیف‌کننده‌های مناسب ترکیبات مولکولی استفاده شد تا همبستگی میان توصیف‌کننده‌ها در نظر گرفته شود. برای ارزیابی روش‌های انتخاب ویژگی نیمه‌نظارتی تُنک، از دو مجموعه داده مولکولی استفاده شد و کارایی این روش‌ها با استفاده از مدل رگرسیون ناپارامتری مبتنی بر هسته گاوسی با روش‌های انتخاب ویژگی نیمه‌نظارتی و غیرنظارتی کلاسیک و انتخاب ویژگی نظارتی تُنک مقایسه شد. نتایج آزمایش‌ها نشان دادند که روش‌های انتخاب ویژگی نیمه‌نظارتی تُنک توانایی بالاتری در انتخاب توصیف‌کننده‌های مناسب برای پیش‌بینی میزان فعالیت ترکیبات مولکولی دارند.

۱۲. مراجع

[1] L. C. Yee and Y. C. Wei, Current modeling methods used in QSAR/QSPR, *Stat. Model. Mol. Descriptors QSAR/QSPR*, vol. 1, (2012).

[2] M. Fernandez, J. Caballero, L. Fernandez, and A. Sarai, Genetic algorithm optimization in drug design QSAR: Bayesian-regularized genetic neural networks (BRGNN) and genetic algorithm-optimized support vectors machines (GA-SVM), *Mol. Divers.*, vol. 15, no. 1, pp. 269–289, (2011).

[3] M. A. Valizade Hasanloei, R. Sheikhpour, M. A. Sarram, *et al.* A combined Fisher and Laplacian score for feature selection in QSAR based drug design using compounds with known and unknown activities. *J Comput Aided Mol Des* 32, 375–384 (2018). https://doi.org/10.1007/s10822-017-0094-6.

[4] S. Uhlmann, S. Kiranyaz, and M. Gabbouj, Semi-supervised learning for ill-posed polarimetric SAR classification, *Remote Sens.*, vol. 6, no. 6, pp. 4801–4830, (2014).

[5] G. Doquire and M. Verleysen, A graph laplacian based approach to semi-supervised feature selection for regression problems, *Neurocomputing*, vol. 121, pp. 5–13, (2013).

[6] Z. Zeng, X. Wang, J. Zhang, Q. Wu, Semi-supervised feature selection based on local discriminative information, Neurocomputing. 173 (2016) 102–109. doi:10.1016/j.neucom.2015.05.119.

[7] X. Song, J. Zhang, Y. Han, J. Jiang, Semi-supervised feature selection via hierarchical regression for web image classification, Multimed. Syst. (2014). doi:10.1007/s00530-014-0390-0.

[8] Y. Han, Y. Yang, Y. Yan, Z. Ma, N. Sebe, S. Member, Semisupervised feature selection via spline regression for video semantic recognition, IEEE Trans. NEURAL NETWORKS Learn. Syst. 26 (2015) 252–264.

[9] M. Fan, X. Zhang, J. Hu, N. Gu, D. Tao, Adaptive Data Structure Regularized Multiclass Discriminative Feature Selection, IEEE Transactions on Neural Networks and Learning Systems. (2021) doi: 10.1109/TNNLS.2021.3071603.

[10] X. Li, Y. Zhang and R. Zhang, Semisupervised Feature Selection via Generalized Uncorrelated Constraint and Manifold Embedding, in *IEEE Transactions on Neural Networks and Learning Systems*. (2021), doi: 10.1109/TNNLS.2021.3069038.

[11] Z. Ma, F. Nie, Y. Yang, J.R.R. Uijlings, N. Sebe, S. Member, et al., Discriminating joint feature analysis for multimedia data understanding, IEEE Trans. Multimed. 14 (2012) 1662–1672.

[12] C. Shi, Q. Ruan, G. An, Sparse feature selection based on graph Laplacian for web image annotation, Image Vis. Comput. 32 (2014) 189–201. doi:10.1016/j.imavis.2013.12.013.

[13] M. Goodarzi, Y. Vander Heyden, S. Funar-Timofei, Towards better understanding of feature-selection or reduction techniques for Quantitative Structure–Activity Relationship models, TrAC Trends Anal. Chem. 42 (2013) 49–63. doi:10.1016/j.trac.2012.09.008.

[14] R. Sheikhpour, M.A. Sarram, S. Gharaghani, M.A. Zare Chahooki, Feature selection based on graph Laplacian by utilizing compounds with known and unknown activities, J. Chemom. (2017). doi:10.1002/cem.2899.

[15] F. Nie, H. Huang, X. Cai, C.H. Ding, Efficient and robust feature selection via joint ℓ2, 1-norms minimization, in: Adv. Neural Inf. Process. Syst., (2010): pp. 1813–1821.

[16] L. Wang, S. Chen, l2,p-matrix norm and its application in feature selection, arXiv Prepr. arXiv1303.3987. (2013).

[17] X. He, D. Cai, P. Niyogi, Laplacian Score for Feature Selection, Adv. Neural Inf. Process. Syst. 18. (2005) 507–514. doi:http://books.nips.cc/papers/files/nips18/NIPS2005_0149.pdf.